To: David Morales, Commissioner, Massachusetts Division of Heath Care Finance and Policy
From: Jonathan Finkelstein, MD, MPH; Richard Platt, MS MD; Jeff Brown, PhD
Department of Population Medicine, Harvard Medical School and Harvard Pilgrim Health Care Institute
Date: May 27, 2010
Re: Public Comment on Proposed Regulation 114.5 CMR 21.00 Massachusetts All-payer Claims Data

Thank you for this opportunity to review and comment on the proposed regulations for creation of an All-payer health plan claims database (APCD). We write as academic health services researchers in the Department of Population Medicine of Harvard Pilgrim Healthcare Institute and Harvard Medical School. We do not represent the views of Harvard Pilgrim Health Care in this matter. We, and our colleagues, are potential users of this data resource for health services, comparative effectiveness, and other types of research designed ultimately to improve the health and health care of the population. Academic researchers in Massachusetts have a long history of using health plan data for high-impact, public-domain research from single insurers as well as collaborative research involving several plans simultaneously. Recently, the NIH's creation of Clinical and Translational Science Centers and the substantial federal investment in comparative effectiveness research have increased the promise of claims-based research.

1. **A distributed analysis model may be superior to aggregation of claims data by the State.** We question whether it is necessary and most effective to aggregate the large amount of patient-level information that is currently reflected in the proposed regulations. We are fully supportive of the need to be able to combine information from residents of the population across insurers, and to compare health care utilization, costs, and outcomes across insurers and care providers. We believe, however, that the use of a distributed model for these types of analyses would be more effective, more reliable, and have less attendant risks than an approach that requires data aggregation in a central repository managed by the State. In the distributed analysis model, each data holder transforms a copy of its data into a standard format, which is then analyzed by programs developed centrally. Advantages of a distributed analysis model include:

- Data holders retain physical control
- Identical programs can work in every location for
  - Data checking
  - Analysis
- Data holders improve the usefulness of their data given their understanding of unique characteristics of their data that are easily lost in aggregation, and can actively engage in data quality checking.
- Better protection of confidential and proprietary data, because users obtain only the information they need
  - Aggregate information is often sufficient
  - Less detailed person-level data is often needed, e.g., data ranges rather than Date of Birth

The experience in the development and use of such models, both locally and nationally, is growing. Investigators in the Department of Population Medicine have developed and participated in a number of networks using these methods, and currently lead a contract with the FDA to develop a data resource that will ultimately have access to data for a population of over 100 million in the US, described below. We attach a recently published paper that describes some of this work, and the data structures that can be developed. This model (developed initially for drug safety studies) does not include certain variable types of interest to the State and researchers (such as cost information), but these could be easily added as additional tables.

2. **The appropriate use of claims data from multiple plans (whether centrally housed or distributed) requires expertise, extensive verification, and development of common definitions.**

Since claims data are designed to pay individual providers, they require substantial data checking and standardization before they can be used for other purposes, including comparisons of utilization and quality assessment. Whether a central or distributed system is ultimately used, sufficient time and effort (both centrally and within each plan) must be committed to resolve these issues. The effort will be greatest in the initial creation of the data source, but will continue to be needed over time as well. It is difficult to independently confirm the completeness of health plan data. However, the types of reliability checks performed (and their results) should be made transparent to the users of the data.

3. **Effective use of the data for academic research or quality measurement will require development of derived data sets using standard definitions, and access through well-designed processes.**

Releasing raw patient-level claims data is likely to result in great variability in the interpretation of data, as well as misinterpretation of the meaning of data fields and values. It will be advantageous to adopt, when suitable, standards and software developed by other multi-payer data systems. This can add value in two ways. First by reducing the time, effort, and cost required to develop data formats, data checking and analysis programs, as well as mechanisms to authorize users for specific uses and access. Second, it will facilitate comparison of Massachusetts' data with other systems'. The FDA has sponsored development of a health plan claims resource, the Mini-Sentinel Distributed Data Network, that is a good candidate for Massachusetts to use as a basis for further development. The Mini-Sentinel uses a distributed data model that does not require submitting any data to FDA or to the FDA's coordinating center. Instead, all analyses are performed by the data holders, using programs provided by the Mini-Sentinel Operations Center, which is based at the Harvard Pilgrim Health Care Institute. The Mini-Sentinel expects to have an operational system covering over 25 million people, including both Harvard Pilgrim and Fallon Community Health Plan members, by the fourth quarter of 2010. Having a Massachusetts all-payer database adopt the FDA Mini-Sentinel's definitions and data structures (whether or not the distributed model was used in creation of the all-payer data resource itself) would allow participation in national studies, and benchmarking with other regions of the US.

4. **Other issues for consideration**

a. Attention to denominators

Health plan data is often thought of as counts of events (ambulatory visits, hospitalizations, procedures). Just as important is the denominator of the eligible person time over which these events occur. Health plan data (and any all-payer data resource) must have accurate data on enrollment/disenrollment for each person (or the ability to limit analyses to patients enrolled for a specific period. This applies to health plan membership overall, but also important changes in benefits (such as major drug coverage changes).

b. Demographic variables

Demographic information on health plan members is generally limited but should include age (to the year for adults and the year and month for children) sex, and race when known. Almost as importantly, a geocode to the census block level should be included, if possible, to allow imputation of sociodemographic variables. The data can include the geocode itself, or be linked to preselected variables so that the location of residence does not remain in the data set itself.

c. Counts

Utilization measured should include inpatient and outpatient utilization, pharmacy dispensings, procedures. Specific rules should be standardized for inclusion of long-term care facilities, rehabilitation services, and mental health services.

d. Cross walks

Patient privacy and compliance with HIPAA are foremost concerns. Clearly, an all-payer claims data set should not include identifying information. However, it is critical that the health plans themselves (who provide the data) are able to retain a cross-walk to patient-level identifying information (name, address, etc.). This is critical to allow studies that can identify patients for further data collection (surveys, collection of specimens) or for intervention (such as care improvement). Obviously, such studies that require contacting patients could only be done with the collaboration of the health plan, approval of relevant institutional review boards for human subjects protection, and reviews for HIPAA compliance.

We believe that the development of an all-payer claims resource in Massachusetts represents a critical opportunity to meet the needs of government, health plans, and the academic community to understand better the care provided to (and required by) the population, its associated costs, and its outcomes. Attention to the above considerations should allow this resource effectively meet these important goals.

Attachment:

Brown JS et al. Distributed health data networks: a practical and preferred approach to multi-institutional evaluations of comparative effectiveness, safety, and quality of care. Medical Care 2010; 48(6 supp):S45-S51.

Distributed Health Data Networks

A Practical and Preferred Approach to Multi-Institutional Evaluations of Comparative Effectiveness, Safety, and Quality of Care

Jeffrey S. Brown, PhD, John H. Holmes, PhD,† Kiran Shah, BA,‡ Ken Hall, MDIV,§ Ross Lazarus, MBBS, MPH,* and Richard Platt, MD, MSc**

Background: Comparative effectiveness research, medical product safety evaluation, and quality measurement will require the ability to use electronic health data held by multiple organizations. There is no consensus about whether to create regional or national combined (eg, "all payer") databases for these purposes, or distributed data networks that leave most Protected Health Information and proprietary data in the possession of the original data holders.

Objectives: Demonstrate functions of a distributed research network that supports research needs and also address data holders concerns about participation. Key design functions included strong local control of data uses and a centralized web-based querying interface.

Research Design: We implemented a pilot distributed research network and evaluated the design considerations, utility for research, and the acceptability to data holders of methods for menu-driven querying. We developed and tested a central, web-based interface with supporting network software. Specific functions assessed include query formation and distribution, query execution and review, and aggregation of results.

Results: This pilot successfully evaluated temporal trends in medication use and diagnoses at 5 separate sites, demonstrating some of the possibilities of using a distributed research network. The pilot demonstrated the potential utility of the design, which addressed the major concerns of both users and data holders. No serious obstacles were identified that would prevent development of a fully functional, scalable network.

Conclusions: Distributed networks are capable of addressing nearly all anticipated uses of routinely collected electronic healthcare data. Distributed networks would obviate the need for centralized databases, thus avoiding numerous obstacles.

Key Words: distributed health data network, all payer databases, comparative effectiveness research network

(*Med Care* 2010;48: S45–S51)

Electronic health records and other data routinely collected during the delivery of, or payment for, health care, and disease or exposure-specific registries are important resources to address the growing need for evidence about the effectiveness, safety, and quality of medical care.[1–6] Unfortunately, even very large individual healthcare databases and registries are not big or diverse enough to address many of needs of clinicians, health care delivery systems, or the public health community. The Institute of Medicine and others have articulated the goal of using routinely collected health information for these "secondary" purposes.[1,7–10] The Federal Coordinating Council for Comparative Effectiveness Research (FCCCER) noted the need for studies "with sufficient power to discern treatment effects and other impacts of interventions among patient subgroups." In their priority recommendations, the FCCCER listed comparative effectiveness research data infrastructure as a primary investment that cuts-across all other comparative effectiveness needs.[5] The FCCCER also listed several considerations for investing in person-level databases for comparative effectiveness research, including the ability to link to external data sources, the research readiness of the databases, and the need to maintain security and privacy of personally identifiable health information.[5]

To address the need to evaluate the processes and outcomes of care of large populations, some propose creation of large, centralized, multipayer claims databases.[11] For example, the Department of Health and Human Services issued a contract titled "Strategic Design for an All-Payor, All-Claims Database to Support Comparative Effectiveness Research." In addition, several states have already implemented

From the *Department of Population Medicine, Harvard Medical School and Harvard Pilgrim Health Care Institute, Boston, MA; †Department of Biostatistics and Epidemiology, Center for Clinical Epidemiology and Biostatistics, University of Pennsylvania School of Medicine, Philadelphia, PA; ‡Lincoln Peak Partners, Westborough, MA; and §Deloitte, NCPHI/OD, Atlanta, GA.

Supported by Agency for Healthcare Research and Quality Contract No. 290–05–0033, US Department of Health and Human Services as part of the Developing Evidence to Inform Decisions about Effectiveness (DEcIDE) program.

The authors of this report are responsible for its content. Statements in this report should not be construed as endorsement by the Agency for Healthcare Research and Quality or the US Department of Health and Human Services.

Reprints: Jeffrey S. Brown, PhD, Department of Population Medicine, Harvard Medical School and Harvard Pilgrim Health Care Institute, 133 Brookline Ave, 6th floor, Boston, MA 02215. E-mail: jeff_brown@harvardpilgrim.org.

Copyright © 2010 by Lippincott Williams & Wilkins

ISSN: 0025-7079/10/4800-0045

Copyright © Lippincott Williams & Wilkins. Unauthorized reproduction of this article is prohibited.

all-payer claims databases to address cost and quality concerns.[12,13] This data centralization approach is alluring because, in theory, it mitigates the complications associated with conducting research across multiple data holders. In practice, a centralized approach raises several serious security, proprietary, operational, legal, and patient privacy concerns for data holders, patients, and funders.[14–17] As one example, even if a centralized database omits explicit identifying information like name and address, it is effectively impossible to prevent reidentification of individual level longitudinal data that contains enough detail to serve multiple purposes. In our experience, these limitations have severely constrained the effective, coordinated use of data held by multiple organizations.

An alternative to centralized, all-payer, databases, is one or more distributed research networks that permit comparative effectiveness and other evaluations across multiple databases without creation of a central data warehouse.[14,15,18–24] Several such networks[18,22–29] currently conduct comparative effectiveness and pharmacoepidemiologic research using a distributed data approach in which data holders maintain control over their protected data and its uses. These networks require data holders to transform their data of interest into a common data model that enforces uniform data element naming conventions, definitions, and data storage formats. The common data format allows data checking, manipulation, and analysis via identical computer programs that are shared by all data holders. Existing distributed networks typically distribute these computer programs via e-mail, data holders manually execute the programs, and return the output via secure e-mail, or another secure mechanism to a coordinating center for aggregation and, possibly, additional analysis. Many studies require no transfer of protected health information. Several single-study networks consisting of 20 to 50 million members also have been developed that adhere to a distributed research network approach.[30,31]

Existing and proposed distributed networks have tremendous potential utility for addressing our current post marketing evidence knowledge gap while benefiting from an approach that is more acceptable to data holders.[14,15,21] In addition, a distributed approach keeps the data close to the people who know the data best and who can best consult on proper use of the data and investigate findings or anomalies.

Obstacles to effective implementation of both centralized and distributed approaches include differences in computing environments and information systems, the need for data standardization and checking, organization-by-organization variation in contracting policies and procedures, concerns related to the ethics of human subjects research and data privacy, and cross-institution variation in the rules and guidelines related to privacy and proprietary issues.[32,33] Distributed networks can be built and tested in phases that allow the network to operate while being built, and network data resources can be updated and enhanced without disrupting overall network operations. Networks have the need for responsible and consistent stewardship of clinical records, and they exchange the requirements of centralized database administrative and computing infrastructure for similar processes on the part of each data holder. Thus, the administrative operation of networks can be cumbersome.

Many of the administrative, technical, and analytic barriers to developing efficient and scalable distributed health data networks to support population-level analyses can be addressed through network features that support the needs of users and data holders. We describe here the design and pilot implementation of a distributed research network infrastructure intended to meet the broad needs of all parties for comparative effectiveness evaluation and other uses. We note the challenges and barriers identified, and provide a blueprint for development of a comprehensive distributed research network as reusable national resource.

METHODS

Background and Needs Assessment

The design, implementation, and evaluation plan of the pilot distributed network was based on findings from previous studies,[14,15,21,34] coupled with our experience in operating a distributed network, the HMO Research Network Center for Education and Research on Therapeutics, and participating in other networks,[23,27,28,31] including a current study of the safety of the H1N1 vaccine.[30] Our prior work investigated the needs of data holders (eg, health plans) and potential network users (eg, federal agencies) with respect to making their data available for comparative effectiveness and other secondary uses.[14,15] Data holders identified several requirements for voluntary participation in a distributed network. These included: (1) complete control of, access to, and uses of, their data, (2) strong security and privacy features, (3) limited impact on internal systems, (4) minimal data transfer, (5) auditable processes, (6) standardization of administrative and regulatory agreements, (7) transparent governance, and (8) ease of participation and use.

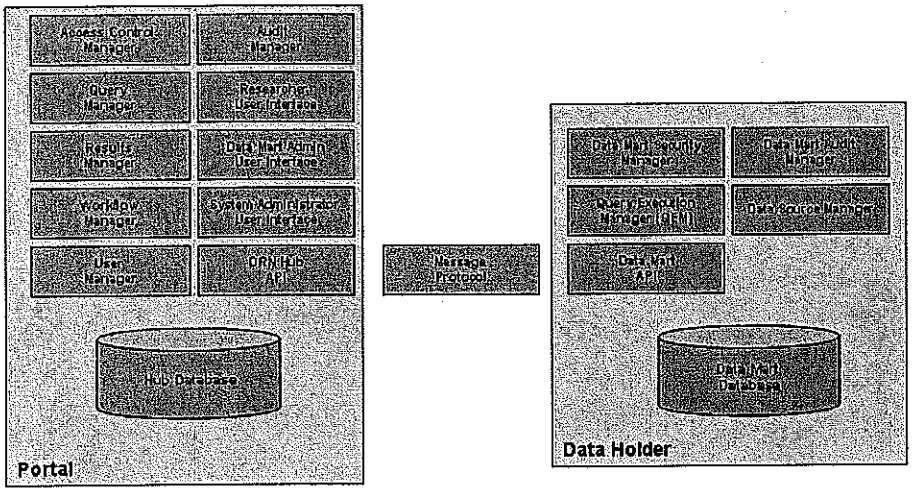
Users' needs assessments, by contrast, did not depend on whether the underlying architecture was a distributed network or centralized database. Potential users identified other key elements of a network, including: menu-driven querying, easy access for feasibility assessments and for public health surveillance and monitoring, the ability to specify and create subsets of the complete data via menu-driven querying or complex programming code, and reuse of network tools to improve efficiency (eg, reuse of validated exposure and outcome algorithms).[15] Nontechnical users wanted the ability to ask simple questions without assistance (eg, counts of people between the ages of 65 and 74 with a positron emission tomography scan in 2008). More sophisticated users wanted the ability to perform complex analyses (eg, compare risk adjusted survival curves for breast cancer patients treated with tamoxifen as adjuvant chemotherapy, to those who were not treated). Potential users also noted that it is often difficult to get rapid responses from existing systems, and this was even noted by users who were also data holders.

Implementation

The design and rapid prototyping process focused on addressing the 8 specific data holder concerns noted above. We used a phased approach to development and implemen-

© 2010 Lippincott Williams & Wilkins

Copyright © Lippincott Williams & Wilkins. Unauthorized reproduction of this article is prohibited.



Note: The demonstration was built using the Microsoft .Net (Microsoft Corporation, Redmond, WA) platform using the C# programming language. The portal-based web application components utilized ASP.NET, with service components arranged in a service-oriented architecture (SOA) utilizing the Simple Object Access Protocol (SOAP) for communications between components. The portal database was built on SQL Server 2005; the query execution manager (QEM) components used ADO.NET and ODBC for connectivity to remote databases.

FIGURE 1. System architecture.

TABLE 1. Description of System Components

Component	Description
Portal	
Access control manager	Manages all aspects of security for portal including authentication, session management, policies, group permissions, user permissions, and access rights.
Query manager	Manages query entry, routing, and distribution.
Results manager	Manages receipt, organization, assembly, merging, and aggregation of result sets.
Workflow manager	Manages workflow (eg, for query approval) including request routing, alerting and notification, approval management, and tracking.
User manager	Manages user accounts.
Audit manager	Provides auditing functions including activity and error logging.
Researcher user interface	User interface for research users, including menu-driven and ad hoc query entry, query management, result status, and result set management.
Data Mart administrator user interface	User interface for Data Mart administrators including data mart setup and configuration, access control management, and workflow management.
System administrator user interface	User interface for System administrators including portal setup and configuration, access control management, and user management.
Hub API	Application Programming Interface (eg, web service API) for the DRN portal. Exposes portal functions for remote applications including query retrieval and results submission.
Hub database	Database for the portal.
Message protocol	Protocol for messaging between the portal and external applications including the Data Marts.
Data holder Data Mart	
Data Mart security manager	Manages all aspects of security for Data Mart including authentication, session management, policies, group permissions, user permissions, and access rights.
Query execution manager	Manages review and execution (or rejection) of queries including queue management, query translation, query engine interface, and results handling.
Data Mart API	Application Programming Interface (eg, web service API) for the Data Mart. Exposes functions for remote applications including query submission and results retrieval.
Data Mart audit manager	Provides auditing functions including activity and error logging.
Data source manager	Manages exchange of data between the Data Mart database and source systems.
Data Mart database	Database for the Data Mart.

tation of the pilot distributed network. The first phase included a web-based portal system for menu-driven querying of summary-level datasets held by 5 data holders (Harvard Pilgrim Health Care; Group Health Cooperative, Geisinger Health Systems, Kaiser Permanente Colorado, and HealthPartners Research Foundation). Figure 1 and Table 1 illustrate and describe the system architecture and features; Figure 2 illustrates the current menu-driven query interface.

© 2010 Lippincott Williams & Wilkins

Copyright © Lippincott Williams & Wilkins. Unauthorized reproduction of this article is prohibited.

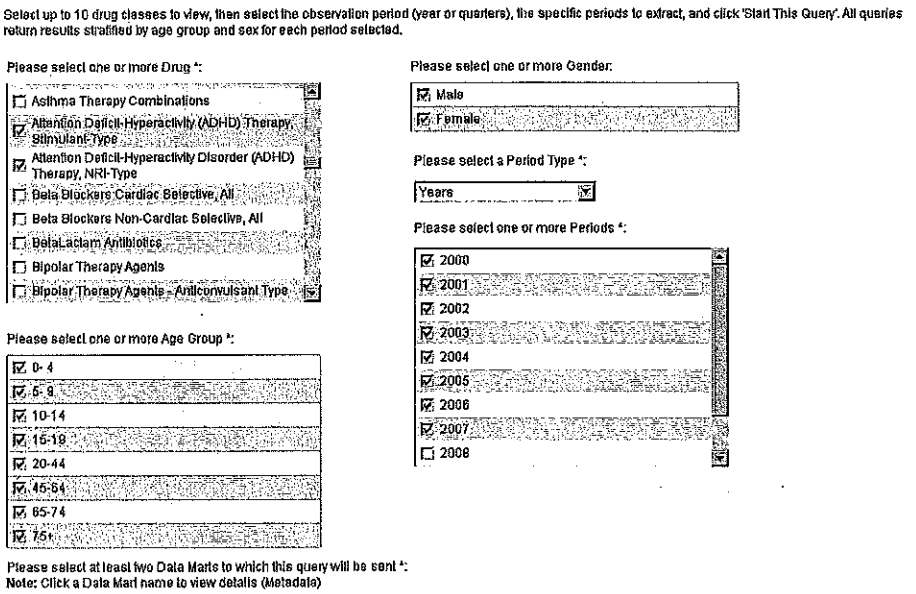


FIGURE 2. Web-based Menu-driven Query Formation. It shows part of the query creation page that allows a user to select specific exposures, diagnoses, or procedures, time periods, and data sources.

Development of the network software relied on a rapid prototyping approach that included multiple rounds of designing, building, testing, and revising of the interface and supporting portal, and the creation of a novel application program, the Query Execution Manager (QEM). Initial querying capability was built for drug utilization by generic name and drug class, diagnosis by 3-digit ICD-9-CM code, and procedure (Healthcare Common Procedure Coding System) queries. Query results were stratifiable by age group, sex, and year or year and quarter. These simple queries were chosen for demonstration purposes, more complex queries are possible within the network design.

From the user perspective, query creation involved logging into the web portal, selecting a query type (eg, generic drug name), selecting query parameters and the data holders to query, and once submitted, reviewing the status of queries and the aggregated query results. Each data holder installed the QEM software and responded to multiple test queries. The system used a "pull" query distribution mechanism that notified (via e-mail) data holders of waiting queries. The data holder user then opened the QEM to review the query details (eg, submitter and reason for submitting) and decides whether to execute, reject, of hold the query for further review. If the data holder decided to run the query, the QEM downloaded the query text from the portal, executed it against a local database, and presented the results to the data holder for review. The data holder could then upload the results to the portal for aggregation with other results and review by the submitter.

Test scenarios were based on summary level data and included assessment of temporal trends in the use of genetic testing, influenza-related medical and pharmacy use, attention-deficit hyperactivity disorder medication use by age, and urticaria diagnoses by age and year, and rate of diabetes by age. We assessed the ability of the system to execute and perform each of the user and data holder tasks described above for each of the test queries submitted. Throughout the rapid prototyping, implementation, and testing process we also informally evaluated data holder acceptance of the system and their willingness to continue development and testing beyond the demonstration. Other network functions such as user-based access control (ie, users have different levels of permissions to submit queries), query formation restrictions (ie, limit on the number and type of query parameters available for selection); and query results viewing rules (ie, requirement of 2 data holder responses before a user can view and export aggregated results) also were developed, tested, and evaluated.

Separately, we partnered with the National Center for Public Health Informatics to design and pilot test an alternative approach for securely transmitting queries and receiving results. The use-case for this pilot test illustrated how an authorized user could securely authenticate to a central portal and securely distribute a computer program to each data holders through their local firewall. The program was executed and the results securely returned for aggregation. Details of this work is described elsewhere.[35]

RESULTS

Implementation and System Functioning

The system was successfully implemented and tested at each of the 5 participating data holders. Each data holder was able to install and operate the QEM, retrieve and execute queries, and upload results. Installation took approximately 15 minutes.

Each of the sample queries was executed without error. Figures 3A and B show how data holders interact with the system, specifically the functions that allow them to review queries before executing them, and review results before uploading them to the central portal. Figure 4 illustrates results from a set of sample queries regarding the use of attention-deficit hyperactivity disorder medications; this type of information can help identify trends that may warrant further evaluation or help

 © 2010 Lippincott Williams & Wilkins

Copyright © Lippincott Williams & Wilkins. Unauthorized reproduction of this article is prohibited.

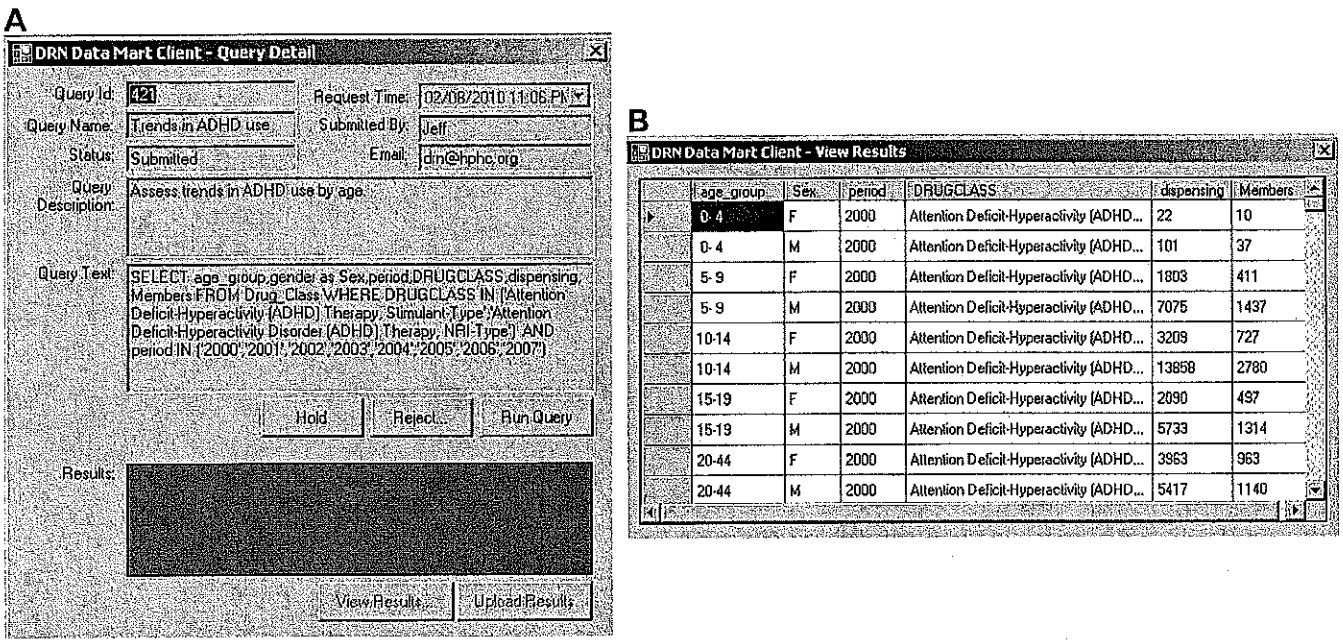


FIGURE 3. Query Execution Manager: Query review and results screen. A, Shows query details that are presented to the data holder for review. In this case the query text is submitted as structured query; alternatively, the design allows a user to send an appropriately structured analytic program. B, Shows local query results for review and approval for upload to the central portal for aggregation.

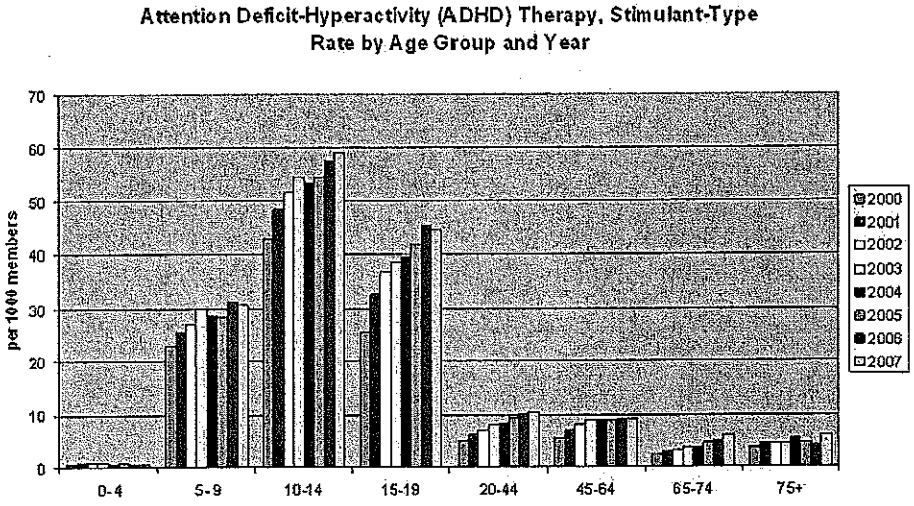


FIGURE 4. Example of findings that could be obtained using the query tool output: Rate of attention-deficit-hyperactivity disorder drug use by age group and year.

understand the adoption and diffusion of new products. The sample queries were completed within a day of submission; this period included local review of incoming queries and of the results.

Evaluation of Acceptability to Data Holders

All data holders found the system acceptable and were willing to continue working on development and implementation. The data holders found the approach acceptable because the design directly addressed many of the data holder concerns listed above. Specifically, the design illustrated: (1) data holder data autonomy that allowed approval of all query executions and data uploads, (2) an easy to use query interface, (3) centralization of the network logic in a portal, and (4) simple installation and light-weight software that did not require any special expertise to install or use. Further, the "pull" mechanism for query distribution (ie, data holders are notified of waiting queries and retrieve them) was also an important favorable factor for data holders' acceptance.

DISCUSSION

Overall, this demonstration validated key design features of a distributed health data network including: use of a

© 2010 Lippincott Williams & Wilkins

Copyright © Lippincott Williams & Wilkins. Unauthorized reproduction of this article is prohibited.

central portal, a "pull" query distribution mechanism that obviated the need to allow queries to pass through firewalls, and local control that allows data holders to maintain physical control of their data and all uses while simultaneously increasing research access for authorized users. These features were paramount to data holders. The use of summary-level data for this demonstration obviated many data privacy concerns, facilitating acceptance by data holders. The technical platform we developed is fully capable of supporting the use of patient and encounter-level utilization information, including the ability to submit a SAS program as the query text.[14]

This design minimizes data holder information technology responsibilities, leaves protected health information under the control of the data holder, provides for a more straightforward security implementation, and focuses network management tasks at the central portal. It supports important capabilities such as secure communications and data protection, auditable processes, a simple query interface, and locally managed query authorization.

The menu-driven interface facilitates the use of the distributed network by users with limited technical expertise. Network features streamline workflow among participating sites and enable an authorized user to quickly assess the feasibility of various comparative effectiveness studies in larger populations than might not normally be readily available. This demonstration project was relatively limited in scope to allow data holders to become comfortable with the technical design and the governance needed to manage and adjudicate simple query requests. More complex and complete demonstrations are currently underway that leverage the same network infrastructure to support full comparative effectiveness studies. Enhancements will allow more sophisticated authorization, security, and permission policies, a more flexible query interface, and access to additional data and query types. In addition, the infrastructure is designed to allow distributed multivariate analyses, either through use the of iterative methods[36,37] or by merging site-specific analysis files that omit PHI, for instance through the use of high dimensionality propensity scores.[38]

Finally, advances in governance are as important to expanding this network model as any of the technical capabilities. Network governance requires policies and procedures to address issues such as data holder protections, conflict of interest, external communications, priority setting, by-laws, data security, accounting, network strategy, stakeholder issues, and HIPAA and human subjects protection. In addition, a coordinating center is needed to maintain network infrastructure, documentation, coordination, monitoring of data resources and contacts, documentation of lessons learned, data validity activities, and study implementation.

CONCLUSION

In theory, either a distributed or a centralized (eg, all-payer) approach can meet the need to use the growing amount of electronic health data to address important societal questions. However, a distributed network is preferred because it can perform essentially all the functions desired of a centralized database, while avoiding many disadvantage of centralized databases. In addition, distributed networks have these advantages, compared with centralized systems: (1) They allow data holders to maintain physical control over their data; without this control, in our experience, data holders are unlikely to voluntarily participate. (2) They ensure ongoing participation of individuals who are knowledgeable about the systems and practices that underlie each data holder's data. (3) They allow data holders to assess and authorize query requests, or categories of requests, on a user-by-user or case-by-case basis. (4) Distributed systems minimize the need to disclose protected health information thus mitigating privacy concerns, many of which are regulated by the Privacy and Security Rules of the Health Insurance Portability and Accountability Act of 1996 (HIPAA). (5) Distributed systems minimize the need to disclose and lose control of proprietary data. (6) A distributed approach eliminates the need to create, secure, maintain, and manage access to a complex central data warehouse. (7) Finally, a distributed network also avoids the need to repeatedly transfer and pool data to maintain a current database, which is a costly undertaking each time updating is necessary.

A phased approach is suggested for implementing a large scale distributed research network infrastructure for comparative effectiveness research and other purposes. Questions that leverage the most commonly used and best understood data types, target large populations, and execute standard statistical analyses using well-developed software packages will be the best candidates for demonstration studies. The Agency for Healthcare Research and Quality's recent grant awards include several examples of research projects that might benefit from the large study populations accessible through distributed data networks, including the study of outcomes resulting from depression treatments and asthma treatments in pregnancy, the study of ACE inhibitors in African-American males, and the study of various treatments for lumbar spine.[39] Similarly, the Food and Drug Administration's planned Sentinel Initiative to monitor the safety of medical products, could also use a distributed data network, either identical to one developed for comparative effectiveness, or very similar to it. Relatively small investments compared with the cost of developing the underlying electronic health information will reduce individual study costs and demonstrate the value of creating reusable, distributed research networks.

REFERENCES

1. Baciu A, Stratton K, Burke SP, eds. *The Future of Drug Safety: Promoting and Protecting the Health of the Public.* Washington, DC: Institute of Medicine of the National Academies; 2006.
2. McClellan M. Drug safety reform at the FDA—pendulum swing or systematic improvement? *N Engl J Med.* 2007;356:1700–1702.
3. Platt R, Wilson M, Chan KA, et al. The new Sentinel Network—improving the evidence of medical-product safety. *N Engl J Med.* 2009;361:645–647.
4. Strom BL. The future of pharmacoepidemiology. In: Strom BL, ed. *Pharmacoepidemiology.* Chichester, United Kingdom: John Wiley & Sons; 2005.
5. Federal Coordinating Council for Comparative Effectiveness Research. Report to the President and Congress. Department of Health and Human Services; 2009. Available at: www.hhs.gov/recovery/prgrams/cer/cerannualrpt.pdf.

© 2010 Lippincott Williams & Wilkins

Copyright © Lippincott Williams & Wilkins. Unauthorized reproduction of this article is prohibited.

6. Gliklich RE, Dreyer NA, eds. Registries for Evaluating Patient Outcomes: A User's Guide. (Prepared by Outcome DEcIDE Center [Outcome Sciences, Inc. dba Outcome] under Contract No. HHSA290200500351 TO1.). AHRQ Publication No. 07-EHC001-1. Rockville, MD: Agency for Healthcare Research and Quality. April 2007.
7. Institute of Medicine. Initial National Priorities for Comparative Effectiveness Research. The National Academies Press: 2009. Available at: http://www.iom.edu/Reports/2009/ComparativeEffectivenessResearchPriorities.aspx.
8. Robert Wood Johnson Foundation. National Effort to Measure and Report on Quality and Cost-Effectiveness of Health Care Unveiled. 2007. Available at: http://www.rwjf.org/pr/product.jsp?id=22371. Accessed February, 2010.
9. Engelberg Center for Health Care Reform at Brookings. Using Data to Support Better Health Care: One Infrastructure with Many Uses. 2009. Available at: http://www.brookings.edu/events/2009/1202_health_care_data.aspx. Accessed February, 2010.
10. Engelberg Center for Health Care Reform at Brookings. Implementing Comparative Effectiveness Research: Priorities, Methods, and Impact. June 2009. Available at: http://www.brookings.edu/~/media/Files/events/2009/0609_health_care_cer/0609_health_care_cer.pdf.
11. Mosquera M. HHS awards $1M contract for effectiveness database. Government Health IT, 2010. Available at: http://govhealthit.com/newsitem.aspx?nid=73035. Accessed January, 2010.
12. Office for Oregon Health Policy and Research. POLICY BRIEF: All-Payer, All-Claims Data Base. 2009. Available at: http://www.oregon.gov/OHPPR/HFB/docs/2009_Legislature_Presentations/Policy_Briefs/PolicyBrief_AllPayerAllClaimsDatabase_4.30.09.pdf.
13. Miller P, Schneider CD. State and National Efforts to Establish All-Payer Claims Databases. Paper presented at: All-payer claims databases: A key to healthcare reform. Massachusetts Health Data Consortium Fall Workshop; 2009; Boston, MA.
14. Brown JS, Holmes J, Maro J, et al. Design specifications for network prototype and cooperative to conduct population-based studies and safety surveillance. Effective Health Care Research Report No. 13. (Prepared by the DEcIDE Centers at the HMO Research Network Center for Education and Research on Therapeutics and the University of Pennsylvania Under Contract No. HHSA290200500331 T05.) Agency for Healthcare Research and Quality: July 2009. Available at: www.effectivehealthcare.ahrq.gov/reports/final.cfm.
15. Maro JC, Platt R, Holmes JH, et al. Design of a national distributed health data network. *Ann Intern Med*. 2009;151:341–344.
16. Rosati K. Using electronic health information for pharmacovigilance: the promise and the pitfalls. *J Health Life Sci Law*. 2009;2:171, 173–239.
17. Rosati KB. HIPAA privacy: the compliance challenges ahead. *J Health Law*. 2002;35:45–82.
18. Hornbrook MC, Hart G, Ellis JL, et al. Building a virtual cancer research organization. *J Natl Cancer Inst Monogr*. 2005:12–25.
19. Lazarus R, Yih K, Platt R. Distributed data processing for public health surveillance. *BMC Public Health*. 2006;6:235.
20. McMurry AJ, Gilbert CA, Reis BY, et al. A self-scaling, distributed information architecture for public health, research, and clinical care. *J Am Med Inform Assoc*. 2007;14:527–533.
21. Moore KM, Duddy A, Braun MM, et al. Potential population-based electronic data sources for rapid pandemic influenza vaccine adverse event detection: a survey of health plans. *Pharmacoepidemiol Drug Saf*. 2008;17:1137–1141.
22. Platt R, Davis R, Finkelstein J, et al. Multicenter epidemiologic and health services research on therapeutics in the HMO Research Network Center for Education and Research on Therapeutics. *Pharmacoepidemiol Drug Saf*. 2001;10:373–377.
23. Wagner EH, Greene SM, Hart G, et al. Building a research consortium of large health systems: the Cancer Research Network. *J Natl Cancer Inst Monogr*. 2005:3–11.
24. Chen RT, Glasser JW, Rhodes PH, et al. Vaccine Safety Datalink project: a new tool for improving vaccine safety monitoring in the United States. The Vaccine Safety Datalink Team. *Pediatrics*. 1997;99:765–773.
25. Vacine Safety Datalink. Available at: http://www.cdc.gov/vaccinesafety/vsd/. Accessed January, 2010.
26. Chan K; HMO Research Network. The HMO Research Network. In: Strom BL, ed. *Pharmacoepediomology*. Chichester, United Kingdom: John Wiley & Sons; 2005.
27. Go AS, Magid DJ, Wells B, et al. The Cardiovascular Research Network: a new paradigm for cardiovascular quality and outcomes research. *Circ Cardiovasc Qual Outcomes*. 2008;1:138–147.
28. Magid DJ, Gurwitz JH, Rumsfeld JS, et al. Creating a research data network for cardiovascular disease: the CVRN. *Expert Rev Cardiovasc Ther*. 2008;6:1043–1045.
29. Davis RL, Kolczak M, Lewis E, et al. Active surveillance of vaccine safety: a system to detect early signs of adverse events. *Epidemiology*. 2005;16:336–341.
30. Federal Immunization Safety Task Force. Federal Plans to Monitor Immunization Safety for the Pandemic 2009 H1N1 Influenza Vaccination Program. 2009. Available at: http://www.flu.gov/professional/federal/fed-plan-to-mon-h1n1-imm-safety.pdf. Accessed January 31, 2010.
31. Velentgas P, Bohn R, Brown JS, et al. A distributed research network model for post-marketing safety studies: the Meningococcal Vaccine Study. *Pharmacoepidemiol Drug Saf*. 2008;17:1226–1234.
32. Greene SM, Geiger AM. A review finds that multicenter studies face substantial challenges but strategies exist to achieve Institutional Review Board approval. *J Clin Epidemiol*. 2006;59:784–790.
33. Greene SM, Geiger AM, Harris EL, et al. Impact of IRB requirements on a multicenter survey of prophylactic mastectomy outcomes. *Ann Epidemiol*. 2006;16:275–278.
34. Brown JS, Moore KM, Braun MM, et al. Active influenza vaccine safety surveillance: potential within a healthcare claims environment. *Med Care*. 2009;47:1251–1257.
35. National Center for Public Health Informatics Public Health Research Grid. Pubic Health Grid Research and Development: DRN Demonstration. 2009. Available at: http://phgrid.blogspot.com/search/label/DRN and http://sites.google.com/site/phgrid/. Accessed February, 2010.
36. Karr A, Lin X, Sanil AP, et al. Secure regression on distributed databases. *J Comput Graph Stat*. 2005;14:263–279.
37. Karr AF, Lin X, Reiter JP, et al. Secure regression on distributed databases. *J Comput Graph Stat*. 2005;14:1–18.
38. Rassen JA, Avorn J, Schneeweiss S. Multivariate-adjusted pharmacoepidemiologic analyses of confidential information pooled from multiple health care utilization databases. *Pharmacoepidemiol Drug Saf*. In press.
39. Agency for Healthcare Research and Quality. Comparative Effectiveness Research Grant Awards. 2010. Available at: http://effectivehealthcare.ahrq.gov/index.cfm/comparative-effectiveness-research-grant-awards/. Accessed January, 2010.

© *2010 Lippincott Williams & Wilkins*

Copyright © Lippincott Williams & Wilkins. Unauthorized reproduction of this article is prohibited.